
A Dual Approach to Imitation Learning from Observations with Suboptimal Offline Datasets

Anonymous Authors

Paper under double-blind review

Abstract

Demonstrations are an effective alternative for instruction specification to learning agents when writing a reward function becomes difficult. However, demonstrating expert behavior in the action space of the agent becomes unwieldy when agents have complex morphologies and when we desire to leverage plethora of internet data (i.e. videos, text) that lack action annotations. In this work, we ask the question if an agent can learn effective behaviors when expert observations are provided to the agent leveraging its own previously collected transitions in the environment. We propose a new method for learning from observations in the fully offline setting by presenting a dual objective that can produce near-optimal imitating policies from the dataset. As opposed to prior approaches in off-policy learning from observations, we do not require learning a discriminator, or an inverse dynamics model and simply learn two function approximators - an actor and a critic. We demonstrate the effectiveness of our approach on a broad range of tasks and datasets from D4RL. Furthermore, for the first time in learning from observations, we show that our objective scales to learning from high-dimensional image observations.

1 Introduction

Imitation Learning (Schaal, 1999) holds the promise of leveraging a few expert demonstrations to train performant agents. This setting is also motivated by literature in behavioral and cognitive sciences (Vogt & Thomaschke, 2007; Byrne & Russon, 1998) where the studies indicate that humans learn by imitation, for instance mimicking other humans or watching tutorial videos. While this is often the motivation, the bulk of imitation learning methods often deal with an impractical setting where the learning agent is allowed to interact with the environment as often as needed. We posit that the main reason humans can imitate fast is due to their knowledge priors of interacting with the environment in the past, distilling motor skills from potentially trying to solve unrelated tasks. Scaling up imitation learning requires us to be able to make effective use of large amounts of internet data available in the form of videos or text. These datasets provide near-expert action-free demonstrations to solve tasks that humans often use in the form of tutorial videos to learn. Our key motivation is to bring imitation learning closer to these practical settings by considering the setup of offline imitation learning from observations, where the agent has access to an offline dataset of action-labeled transitions that may be of arbitrary quality collected by attempting unrelated tasks and is provided with a few task-relevant expert observations.

Imitation Learning from Observations (LfO) has been widely studied in the online setting, where the agent is allowed to interact with the environment. Contrary to the setting where expert actions are given along with observations, learning from observations alone is provably more challenging (Kidambi et al., 2021). Some prior works (Torabi et al., 2018a; Edwards et al., 2019) have approached this problem by learning an Inverse Dynamics Model (IDM) by interactions with the environment and then resorting to traditional imitation learning algorithms by first inferring the expert actions from the inverse dynamics models. Distribution matching has served as a principled approach to imitation learning (Ghasemipour et al., 2020). For this class of methods, the learning agent minimizes the

divergence between its own visitation distribution to expert’s visitation distribution. Prior works based on distribution matching either require on-policy interactions (Ho & Ermon, 2016; Torabi et al., 2018b; Zhu et al., 2020; Hoshino et al., 2022), or learn a discriminator and treat the discriminator as a pseudo reward in the offline setting (Ma et al., 2022), require modeling density ratios (Ni et al., 2021; Lee et al., 2019), or optimize a loose upper bound of the true objective to derive an off-policy objective (Zhu et al., 2020; Ma et al., 2022). RL on top of a learned discriminator as a pseudoreward is known to result in compounding errors (Sikchi et al., 2023a) leading to poor downstream policies. In this work, we seek to bypass learning a discriminator or IDM altogether and derive an objective that solves distribution matching exactly without resorting to an upper bound. The key question we seek to answer in this work is — *Can we derive an efficient, lightweight yet principled off-policy algorithm for learning from observations from datasets of arbitrary quality?*

In this work, we propose Dual Imitation Learning from Observations or DIL0, an efficient off-policy algorithm for learning from observations. DIL0 is based on a novel distribution matching objective that admits a convex dual that is easy to optimize. Notably the convex dual learns the optimal value functions of the form $Q^*(s, s')$, by only requiring to sample multiple consecutive states from the dataset. Contrary to existing off-policy methods for LfO (Ma et al., 2022; Zhu et al., 2020), DIL0 optimizes for the exact distribution matching objective without resorting to a upper bound. Unlike prior methods, DIL0 solves a single-player objective making the learning stable and more performant and does not suffer the compounding error issues persistent in methods learning from pseudoreward in the form of discriminator. Our experimental evaluation on a suite of MuJoCo (Todorov et al., 2012) environments and offline datasets from D4RL (Fu et al., 2020) show that DIL0 achieves improved performance consistently over the evaluation suite. Furthermore, we demonstrate that DIL0 scales to image observations seamlessly without extensive hyperparameter tuning.

2 Related Work

Imitation Learning: Imitation Learning deals with the problem of mimicking expert behavior from a limited set of expert demonstrations (state-action pairs). Prior approaches to imitation learning can be understood through the lens of Apprenticeship Learning/Inverse Reinforcement Learning or Distribution Matching. IRL approaches (Abbeel & Ng, 2004) can be derived by considering a min-max game between a policy agent and a reward agent, where the policy agent maximizes the reward function and the reward agent chooses a reward function that maximizes the performance gap between the policy agent and the expert (Sikchi et al., 2022). Distribution Matching approaches like GAIL (Ho & Ermon, 2016), AIRL (Fu et al., 2017), f -MAX (Ghasemipour et al., 2020) learn a policy whose induced visitation distribution matches the visitation distribution of the expert (Ghasemipour et al., 2020). There are deep connections between both the approaches above, either through the lens of duality (Garg et al., 2021; Sikchi et al., 2023b), or by treating them as different strategies to solve a two-player game (Swamy et al., 2021). While the discussed approaches above learn an imitating policy that avoids quadratic regret on the horizon, a simple yet popular and successful alternative to imitation learning is Behavior Cloning (Pomerleau, 1991). Behavior Cloning has seen numerous successes in scaling up with data, partly due to the supervised nature of its objective making it easier to optimize, and partly due to the fact that large amounts of data inherently reduce regret.

Learning from Observations: Learning from Observations (LfO) assumes that expert actions are unavailable. This setting is more practical as performant algorithms developed for LfO can unlock learning from a plethora of internet videos and develop ways to transfer skills across embodiments. Current methods in LfO are often restricted to small observation dimensions and involve complicated learning algorithms. Behavior Cloning is no longer an alternative here as there are no expert actions to clone. Prior works have attempted to infer expert actions by learning an inverse dynamics model (Torabi et al., 2018a; Yang et al., 2019; Edwards et al., 2019) and are based on the assumption of injective dynamics. Errors arising through one-step inverse dynamics model will compound when the policy is deployed often leading to poor performance (Sikchi et al., 2022). Distribution matching approaches to LfO have considered state-next state (Torabi et al., 2019; Zhu et al., 2020; Sun et al.,

2019) visitation distribution matching or purely state-marginal matching (Ni et al., 2021). While these methods have been successful in the online setting where an agent can interact with the environment, in the offline setting deriving an update rule for distribution matching required optimizing an upper bound (Zhu et al., 2020) or learning a discriminator and using it as a psuedoreward for downstream RL (Ma et al., 2022). Discriminators can easily overfit when there are few-expert demonstrations or when learning from high-dimensional observations. The errors in the discriminator theoretically and empirically compound during downstream RL (Sikchi et al., 2023a).

3 Preliminaries

We consider a learning agent in a Markov Decision Process (MDP) (Puterman, 2014; Sutton & Barto, 2018) which is defined as a tuple: $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma, d_0)$ where \mathcal{S} and \mathcal{A} denote the state and action spaces respectively, P denotes the transition function with $P(s'|s, a)$ indicating the probability of transitioning from s to s' taking action a ; R denotes the reward function and $\gamma \in (0, 1)$ specifies the discount factor. The reinforcement learning objective is to obtain a policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ that maximizes expected return: $\mathbb{E}_\pi [\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$, where we use \mathbb{E}_π to denote the expectation under the distribution induced by $a_t \sim \pi(\cdot|s_t)$, $s_{t+1} \sim p(\cdot|s_t, a_t)$ and $\Delta(\mathcal{A})$ denotes a probability simplex supported over \mathcal{A} . f -divergences define a measure of distance between two probability distributions given by $D_f(P||Q) = \mathbb{E}_{x \sim Q} [f(\frac{P(x)}{Q(x)})]$ where f is a convex function.

Visitation distributions and Dual RL: A key quantity of interest in this work is the visitation distribution of a policy denoted by $d^\pi(s, a)$. The visitation distribution is defined as the discounted probability of visiting a particular state under policy π , i.e $d^\pi(s, a) = (1 - \gamma)\pi(a|s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\pi)$ and uniquely characterizes the policy π that achieves the visitation distribution as follows: $\pi(a|s) = \frac{d^\pi(s, a)}{\sum_a d^\pi(s, a)}$. Dual-RL and DICE approaches (Sikchi et al., 2023b; Nachum & Dai, 2020) frames RL treating visitation distributions as primary objects of interest. The primal objective for RL when formulated in terms of visitation distributions is given by:

$$\begin{aligned} & \max_{d \geq 0} \mathbb{E}_{d(s, a)} [r(s, a)] - \alpha D_f(d(s, a) || d^O(s, a)) \\ \text{s.t } & \sum_{a \in \mathcal{A}} d(s, a) = (1 - \gamma)d_0(s) + \gamma \sum_{(s', a') \in \mathcal{S} \times \mathcal{A}} d(s', a') p(s|s', a'), \forall s \in \mathcal{S}. \end{aligned} \quad (1)$$

The above objective is constrained and difficult to optimize, but the Lagrangian dual of the above objective presents an unconstrained optimization that results in a performant Dual-RL algorithm.

$$\min_V (1 - \gamma) \mathbb{E}_{s \sim d_0} [V(s)] + \alpha \mathbb{E}_{(s, a) \sim d^O} [f_p^* ([TV(s, a) - V(s)] / \alpha)], \quad (2)$$

where $f_p^*(y) = \max_{x \in \mathbb{R}} \langle x \cdot y \rangle - f(x)$ s.t $x \geq 0$.

Imitation Learning from Observations: Imitation learning considers the setting where an agent is provided with expert demonstrations and lacks access to a reward function. The goal of imitation learning is to find a policy that mimics expert behavior. Learning from Demonstrations, or LfD considers the setting where the expert provides state-action trajectories: $\mathcal{D}^E = \{[s_0^0, a_0^0, s_1^0, a_1^0, \dots, s_h^0, a_h^0], \dots, [s_0^n, a_0^n, s_1^n, a_1^n, \dots, s_h^n, a_h^n]\}$. Learning from observations, or LfO presents a more challenging objective where the expert provides observation-only trajectories: $\mathcal{D}^E = \{[s_0^0, s_1^0, \dots, s_h^0], \dots, [s_0^n, s_1^n, \dots, s_h^n]\}$. Prior works have demonstrated that learning from demonstrations can be formulated as a distribution matching problem (Ghasemipour et al., 2020), minimizing the divergence between expert visitation $d^E(s, a)$ and agent visitation distribution $d^\pi(s, a)$. Similarly, in the setting of learning from observations prior works have proposed distribution matching objectives that either match state-marginal visitations (Ni et al., 2021) or {state,next-state} visitation distributions (Torabi et al., 2018b).

Our work focuses on the offline setting where we have access to a dataset of transitions of arbitrary quality the agent might have collected in the past. The expert demonstrates observation-only trajectories and the objective is to learn a policy that minimizes the divergence with expert visitation

solely using the offline dataset and expert observations. We denote the offline dataset by d^O consisting of $\{\text{state}, \text{action}, \text{next-state}\}$ tuples and $\rho(s, a, s')$ as the corresponding visitation distribution of offline dataset.

4 DILO: Dual Imitation Learning from Observations

We propose DILO, an approach that directly infers expert value function from expert observation trajectories using offline datasets. This objective allows us to infer $Q^*(s, s')$ which is later used to distill an optimal policy. We extend the work of Sikchi et al. (2023b) which first proposed mixture distribution matching as a way to leverage offline data in the context of imitation learning by proposing a novel method to *replace their action-dependent objective with an action-free objective*.

In the setting of imitation learning from observations, the demonstrated data lacks expert actions. A key observation, also leveraged by some prior works (Torabi et al., 2018b), is that the next-state encodes the information about missing expert actions as the next-state is a stochastic function of the current state and action. To leverage this insight in the setting of distribution matching, we define $\{\text{state}, \text{next-state}, \text{next-action}\}$ visitation distributions denoted by $\tilde{d}^\pi(s, s', a')$. Formally, it extends the definition of state-action visitation distribution by denoting the discounted probability of reaching the $\{\text{state}, \text{next-state}\}$ pair s, s' under policy π and subsequently taking an action a' :

$$\tilde{d}(s, s', a') = (1 - \gamma)\pi(a'|s') \sum_{s_0 \sim d_0, a_t \sim \pi(s_t)} \gamma^t p(s_{t+1} = s', s_t = s | \pi) \quad (3)$$

To derive our method, we first consider a known expert $\{\text{state}, \text{next-state}, \text{next-action}\}$ visitation distribution. We will subsequently introduce a principled way to remove the need to sample an expert action. Let $\tilde{d}^E(s, s', a')$ denote the expert state-action visitation distribution, and $\rho(s, s', a')$ denote the offline visitation distribution. We first consider a mixture distribution matching objective for imitation learning that allows us to incorporate the offline data in the imitation learning objective:

$$\min_{\pi} \mathcal{D}_f(\text{Mix}_{\beta}(\tilde{d}^\pi(s, s', a'), \rho(s, s', a')) \| \text{Mix}_{\beta}(\tilde{d}^E(s, s', a'), \rho(s, s', a'))), \quad (4)$$

where for any two distributions μ_1 and μ_2 , $\text{Mix}_{\beta}(\mu_1, \mu_2)$ denotes the mixture distribution with coefficient $\beta \in (0, 1]$ defined as $\text{Mix}_{\beta}(\mu_1, \mu_2) = \beta\mu_1 + (1 - \beta)\mu_2$. The above mixture distribution objective is a principled objective for imitation learning — at convergence, $\tilde{d}^\pi(s, s', a') = \tilde{d}^E(s, s', a')$ holds, which implies $d^\pi(s, s') = d^E(s, s')$ and also $d^\pi(s) = d^E(s)$, thus providing an exact reduction to the imitation learning problem formulated as distribution matching (Ghasemipour et al., 2020). Prior work (Kostrikov et al., 2019; Sikchi et al., 2023b;a) has also shown that mixture distribution matching in the space of state-action visitation presents an effective strategy, both theoretically and practically, to leverage offline datasets to learn performant policies. The mixture distribution matching objective can be rewritten as a convex program with linear constraints:

$$\begin{aligned} & \max_{\tilde{d} \geq 0} -\mathcal{D}_f(\text{Mix}_{\beta}(\tilde{d}, \rho) \| \text{Mix}_{\beta}(\tilde{d}^E, \rho)) \\ \text{s.t. } & \sum_{a''} \tilde{d}(s', s'', a'') = (1 - \gamma)\tilde{d}_0(s', s'') + \gamma \sum_{s, a' \in \mathcal{S} \times \mathcal{A}} \tilde{d}(s, s', a') p(s'' | s', a'), \quad \forall s', s'' \in \mathcal{S} \times \mathcal{S}, \end{aligned} \quad (5)$$

where the constraints above dictate the constraints that any valid visitation distribution $\tilde{d}(s', s'')$ needs to satisfy and is a modification to the commonly known *bellman flow constraints*. Our core insight and departure from prior work Sikchi et al. (2023b) is imposing constraints on action-marginalized $\{\text{state}, \text{next-state}, \text{next-action}\}$ visitation distributions that allow us to derive an *action-free* objective for LfO.

Our proposed objective, DILO, is obtained by taking the Lagrangian dual of the above primal objective in Eq 5 given as follows:

$$\begin{aligned} \text{DILO: } \min_Q & \beta(1 - \gamma)\mathbb{E}_{\tilde{d}_0}[Q(s, s')] + \mathbb{E}_{s, s', a' \sim \text{Mix}_{\beta}(\tilde{d}^E, \rho)}[f_p^*(\gamma\mathbb{E}_{s'' \sim p(\cdot | s', a')}[Q(s', s'')] - Q(s, s'))] \\ & - (1 - \beta)\mathbb{E}_{s, s', a' \sim \rho}[\gamma\mathbb{E}_{s'' \sim p(\cdot | s', a')}[Q(s', s'')] - Q(s, s')], \end{aligned} \quad (6)$$

where f_p^* is a variant of conjugate f^* defined as $f_p^*(x) = \max(0, f'^{-1}(x))(x) - f(\max(0, f'^{-1}(x)))$ and Q is the Lagrange dual variable defined as $Q : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}$. Theorem 4.1 below shows that strong duality holds and the dual objective admits the same solution with the added benefit of being action-free. The solution to the dual objective in Equation 6, $Q^*(s, s')$ represents the discounted utility of transitioning to a state s' from s under the optimal imitating policy that minimizes the f -divergence with the expert visitation, or equivalently represents the expected return of optimal policy under a policy dependent reward function $r(s, a) = -\frac{d^{\pi^*}(s, a)}{\rho(s, a)} f(\frac{d^{\pi^*}(s, a)}{\rho(s, a)})$.

Theorem 4.1. *The dual problem to the primal occupancy matching objective (Equation 5) is given by the DILO objective in Equation 6. Moreover, as strong duality holds from Slater’s conditions the primal and dual share the same optimal solution d^* for any offline transition distribution ρ and any choice of mixture distribution ratio β .*

Proof. Proof is deferred to Appendix 6.1.1 due to space constraints. \square

An empirical estimator for the DILO objective in Eq 6 only requires sampling s, s', s'' under a mixture offline dataset and expert dataset and no longer requires knowing any of the actions that induced those transitions. This establishes DILO as a principled action-free alternative to optimizing the occupancy matching objective for offline settings.

4.1 Policy Extraction and Practical Algorithm

To instantiate our algorithm, we use the Pearson Chi-square divergence ($f(x) = (x - 1)^2$) which has been found to lead to stable DICE and Dual-RL algorithms in the past (Al-Hafez et al., 2023). At convergence, the DILO objective does not directly give us the optimal policy π^* but rather provides us with a utility function $Q^*(s, s')$ that quantifies the utility of transitioning to state s' from s in minimizing the f -divergence with the expert’s visitation. To recover the policy we use value-weighted regression which has been shown to provably maximize the Q function while subject to distribution constraint.

$$\mathcal{L}(\psi) = \mathbb{E}_{s, a, s' \sim \rho} \left[e^{\tau Q^*(s, s')} \pi_\psi(a|s) \right] \quad (7)$$

Choice of $\tilde{d}_0(s, s')$: A distribution over state and next-state is implicitly dependent on the policy that induces the next-state. This initial distribution in Eq 6 forms the distribution over states from which the learned policy will acquire effective imitation behavior to mimic the expert. In our work, we set $\tilde{d}_0(s, s')$ to be the uniform distribution over replay buffer $\{s, s'\}$ pairs ensuring that the learned policy is robust to imitate from any starting transition observed from all the transitions available to us.

Practical optimization difficulty of dual objectives:

Prior works in reinforcement learning that have leveraged a dual objective based on bellman-flow constraints suffer from learning instabilities if gradient descent is used directly. Intuitively, in our case learning instability arises as the gradients from $Q(s, s')$ and $Q(s', s'')$ can conflict if the network learns similar feature representations for nearby states due to feature co-adaptation (Kumar et al., 2021). Prior works (Sikchi et al., 2023b) have resorted to using semi-gradient approaches but do not converge provably to the optimal solution (Mao et al., 2024). To sidestep this issue, we leverage the orthogonal gradient update proposed by ODICE (Mao et al., 2024) for the offline RL setting that fixes the conflicting gradient by combining the projection of the gradient of $Q(s', s'')$ on $Q(s, s')$ and the orthogonal component in

Algorithm 1: DILO

- 1: Init Q_ϕ, π_ψ
 - 2: Params: temperature τ , mixture ratio β
 - 3: Let $\mathcal{D} = \hat{\rho} = \{(s, a, s')\}$ be an offline dataset and $\mathcal{D}^\mathcal{E} = \{s, s'\}$ be expert demonstrations dataset.
 - 4: **for** $t = 1..T$ iterations **do**
 - 5: Train Q_ϕ via Orthogonal gradient update on Eq. 6
 - 6: Update π_ψ via Eq. 7
 - 7: **end for**
-

a principled manner. We refer to the ODICE work for detailed exposition. Our complete practical algorithm can be found in Algorithm 1.

5 Experiments

In our experiments, we aim to understand where the prior LfO methods based on IDM or a discriminator fail and how the performance of DILo compares to baselines under a diverse set of datasets. Our experiments with proprioceptive observations consider an extensive set of 24 datasets with varying duality. The simplicity of DILo also motivates us to investigate learning from observation in image-observation based domains.

5.1 Offline Imitation from Observation Benchmarking

To evaluate the efficacy of our proposed method, we consider an extensive offline benchmark comprising locomotion and manipulation tasks where the datasets are constructed with D4RL datasets (Fu et al., 2020) generated in MuJoCo simulator (Todorov et al., 2012). We use the offline imitation benchmark from (Sikchi et al., 2023b) which itself is an extended benchmark from (Ma et al., 2022). The locomotion tasks of the dataset comprise of 1-million transitions from random or medium datasets mixed with 200 expert trajectories. The few-expert setting consists of 30 expert trajectories. In both settings, the expert only provides 1 observation-only expert trajectory. For manipulation environments, we have suboptimal datasets comprising of 30 expert trajectories mixed with human or cloned datasets from D4RL.

Suboptimal Dataset	Env	Access to expert actions					No expert actions			Expert
		RCE	BC (only expert data)	BC (full dataset)	IQ-Learn (offline)	ReCOIL	ORIL	SMODICE	DILo	
random+ expert	hopper	51.41±38.63	4.52±1.42	5.64±4.83	1.85 ± 2.19	108.18±3.28	73.93±11.06	101.61±7.69	97.87±8.11	111.33
	halfcheetah	64.19±11.06	2.2±0.01	2.25±0.00	4.83±7.99	80.20±6.61	60.49±3.53	80.16±7.30	91.18±0.24	88.83
	walker2d	20.90±26.80	0.86±0.61	0.91±0.5	0.57±0.09	102.16±7.19	2.86±3.39	105.86±3.47	108.41±0.64	106.92
	ant	105.38±14.15	5.17±5.43	30.66±1.35	42.23±20.05	126.74±4.63	73.67±12.69	126.78±5.12	122.15±5.15	130.75
random+ few-expert	hopper	25.31±18.97	4.84±3.83	3.0±0.54	1.37 ± 1.23	97.85±17.89	42.04±13.76	60.11±18.28	93.73±7.59	111.33
	halfcheetah	2.99±1.07	-0.93±0.35	2.24±0.01	1.14±1.94	76.92±7.53	2.84±5.52	2.28±0.62	52.32±10.72	88.83
	walker2d	40.49±26.52	0.98±0.83	0.74±0.20	0.39±0.27	83.23±19.00	3.22±3.29	107.18±1.87	108.42±0.25	106.92
	ant	67.62±15.81	0.91±3.93	35.38±2.66	32.99±3.12	67.14± 8.30	25.41 ± 8.58	-6.10±7.85	117.50±4.75	130.75
medium+ expert	hopper	58.71±34.06	16.09±12.80	59.25±3.71	12.90±24.00	88.51±16.73	61.68±7.61	49.74±3.62	99.97±12.62	111.33
	halfcheetah	65.14±13.82	-1.79±0.22	42.45± 0.42	25.67±20.82	81.15±2.84	54.66±0.88	59.50±0.82	90.47±0.64	88.83
	walker2d	96.24±14.04	2.43±1.82	72.76±3.82	59.37±30.14	108.54±1.81	8.19±7.70	2.62±0.93	77.16±6.96	106.92
	ant	86.14±38.59	0.86±7.42	95.47±10.37	37.17±41.15	120.36±7.67	102.74±6.63	104.95±6.43	102.89±3.57	130.75
medium few-expert	hopper	66.15±35.16	7.37±1.13	46.87±5.31	11.05±20.59	50.01±10.36	17.40±15.15	47.61±7.08	41.80±14.81	111.33
	halfcheetah	61.14±18.31	-1.15±0.06	42.21±0.06	26.27±20.24	75.96±4.54	43.24±0.75	46.45±3.12	74.71±6.35	88.83
	walker2d	85.28±34.90	2.02±0.72	70.42±2.86	73.30±2.85	91.25±17.63	6.81±6.76	6.00±6.69	66.64±6.05	106.92
	ant	67.95±36.78	-10.45±1.63	81.63±6.67	35.12±50.56	110.38±10.96	81.53±8.61	80.49±7.69	88.03±9.01	130.75
cloned+expert	pen	19.60±11.40	13.95±11.04	34.94±11.10	2.18±8.75	95.04±4.48	-3.10±0.40	-3.36±0.71	101.36±3.48	106.42
	door	0.08± 0.15	-0.22±0.05	0.01±0.00	0.07±0.02	102.75±4.05	-0.33±0.01	0.25± 0.54	105.60±0.28	103.94
	hammer	1.95±3.89	2.41±4.48	5.45± 7.84	0.27±0.02	95.77±17.90	0.25± 0.01	0.15± 0.078	112.55±22.10	125.71
human+expert	pen	17.81±5.91	13.83±10.76	90.76±25.09	14.29±28.82	103.72±2.90	-3.38±2.29	-2.20±2.40	88.61±14.30	106.42
	door	-0.05±0.05	-0.03±0.05	103.71±1.22	5.6±7.29	104.70±0.55	-0.33±0.01	-0.20± 0.11	101.51±0.99	103.94
	hammer	5.00±5.64	0.18±0.14	122.61±4.85	5.32±1.38	125.19±3.29	1.89±0.70	-0.07±0.39	117.52±8.55	125.71
partial+expert	kitchen	6.875±9.24	2.5±5.0	45.5±1.87	0.0±0.0	60.0±5.70	0.00±0.00	39.16± 1.17	43.0±5.30	75.0
mixed+expert	kitchen	1.66±2.35	2.2±3.8	42.1±1.12	0.0±0.0	52.0±1.0	0.00±0.00	42.5±2.04	44.0±8.30	75.0

Table 1: The normalized return obtained by different offline IL (both provided with and without expert actions) methods trained on the D4RL suboptimal datasets with 1 expert trajectory. The statistics are obtained over 5 random seeds. Methods with avg. perf within the std-dev of the top performing method is highlighted.

Baselines: We compare DILo against offline imitation from observations (LfO) methods such as ORIL (Zolna et al., 2020), SMODICE (Ma et al., 2022) as well as offline imitation from demonstration methods like BC (Pomerleau, 1991), IQ-Learn (Garg et al., 2021) and ReCOIL (Sikchi et al., 2023b). We choose these imitation learning methods as they represent the frontier of LfO and LfD setting outperforming methods like ValueDICE (Kostrikov et al., 2019) and DemoDICE (Kim et al., 2022) as shown in prior works. Intuitively, the imitation from demonstration results represent the upper bound of performance as they have additional information on expert actions even though sometimes we observe LfO algorithms to surpass them in performance. ORIL and SMODICE require learning a discriminator and subsequently a run downstream RL treating the discriminator as a pseudo-reward. Errors in learned discriminator can lead to compounding errors in the downstream RL step. We observe this effect in the datasets where we have few expert trajectories and the discriminator overfits.





					
		Lift-MG	Lift-MH	Can-MG	Can-MH
State 50 Demos	BCO	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0	0.0 \pm 0.0
	SMODICE	0.41 \pm 0.02	0.46 \pm 0.1	0.54 \pm 0.01	0.28 \pm 0.01
	DILO	0.59 \pm 0.03	0.97 \pm 0.02	0.53 \pm 0.02	0.64 \pm 0.03
Image 50 Demos	BCO	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00
	SMODICE	0.21 \pm 0.02	0.4 \pm 0.12	0.1 \pm 0.04	0.02 \pm 0.01
	DILO	0.76 \pm 0.08	0.94 \pm 0.02	0.25 \pm 0.02	0.15 \pm 0.01

Table 2: Side-by-side comparison of LfO methods on state-only imitation vs image-only imitation. **DILO** shows noticeable improvement over existing LfO methods without hyperparameter tuning. Columns denote different suboptimal datasets (MG: Machine Generated, MH: Multi-Human)

DILO outperforms all LfO baselines across most datasets. In high-dimensional tasks like dextrous manipulation prior methods completely fail likely due to discriminator overfitting while **DILO** gets rid of this intermediate step completely rather reducing the problem of LfO to a similar training setup as an traditional actor-critic algorithm.

5.2 Imitating from Expert Image Observations

Imitation from Image Observations presents a challenging but realistic problem-setting for the agents to be able to learn from internet scale videos. To evaluate our algorithm, we consider the Robomimic datasets (Mandlekar et al., 2021) which provide suboptimal datasets along with expert datasets. Our suboptimal datasets comprises of (Multi-Human (MH), Machine Generated (MG)) datasets without access to expert trajectories. We obtain 50 expert-observation trajectories from Proficient Human (PH) datasets. This task is more complicated as the agent has to learn expert actions purely from OOD datasets and match expert visitations. In this setup we consider the most performant LfO baseline from the previous section SMODICE (Ma et al., 2022) along with a BCO (Torabi et al., 2018a) baseline which runs behavior cloning by inferring expert actions using the suboptimal dataset since BC approaches have seen numerous successes in scaling up with image observations.

Table 2 shows the result of these approaches on 4-different datasets using both state and image observations. SMODICE shows competitive results when learning from state-observations but does not scale up well to images likely due to the overfitting of the discriminator in high-dimensional space. BCO fails consistently across both state and image experiments. **DILO** outperforms baselines and demonstrates improved performance across both state and image observations.

6 Conclusion

Learning from Observations (LfO) holds the potential to use large scale internet data (videos, texts, etc) as a task specification. With the agent’s own prior interaction with the world, the cross-embodiment gap between expert-data and offline-data as the agent can learn actions that induce the same visitation distribution as expert. In this work, we present **DILO**, a dual approach to imitation learning from observation that learns from expert observations and offline suboptimal agent interaction data. Unlike prior approaches, **DILO** avoids the need to learn a discriminator or inverse dynamics model common in prior work. These favorable properties reduce LfO to the same complexity as a traditional actor-critic RL algorithm, thus paving the way to scale up to image observations. Our experiments demonstrate strong performance in a broad range of locomotion and manipulation tasks with varying quality of offline datasets.

References

- Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.
- Firas Al-Hafez, Davide Tateo, Oleg Arenz, Guoping Zhao, and Jan Peters. Ls-ic: Implicit reward regularization for inverse reinforcement learning. *arXiv preprint arXiv:2303.00599*, 2023.
- Richard W Byrne and Anne E Russon. Learning by imitation: A hierarchical approach. *Behavioral and brain sciences*, 21(5):667–684, 1998.
- Ashley Edwards, Himanshu Sahni, Yannick Schroecker, and Charles Isbell. Imitating latent policies from observation. In *International Conference on Machine Learning*, pp. 1755–1763. PMLR, 2019.
- Ben Eysenbach, Sergey Levine, and Russ R Salakhutdinov. Replacing rewards with examples: Example-based policy search via recursive classification. *Advances in Neural Information Processing Systems*, 34:11541–11552, 2021.
- Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.
- Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.
- Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.
- Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon. Iq-learn: Inverse soft-q learning for imitation. *Advances in Neural Information Processing Systems*, 34:4028–4039, 2021.
- Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, pp. 1259–1277. PMLR, 2020.
- Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29:4565–4573, 2016.
- Hana Hoshino, Kei Ota, Asako Kanezaki, and Rio Yokota. Opirl: Sample efficient off-policy inverse reinforcement learning via distribution matching. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 448–454. IEEE, 2022.
- Rahul Kidambi, Jonathan Chang, and Wen Sun. Mobile: Model-based imitation learning from observation alone. *Advances in Neural Information Processing Systems*, 34, 2021.
- Geon-Hyeong Kim, Seokin Seo, Jongmin Lee, Wonseok Jeon, HyeongJoo Hwang, Hongseok Yang, and Kee-Eung Kim. Demodice: Offline imitation learning with supplementary imperfect demonstrations. In *International Conference on Learning Representations*, 2022.
- Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. *arXiv preprint arXiv:1912.05032*, 2019.
- Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.
- Aviral Kumar, Rishabh Agarwal, Tengyu Ma, Aaron Courville, George Tucker, and Sergey Levine. Dr3: Value-based deep reinforcement learning requires explicit regularization. *arXiv preprint arXiv:2112.04716*, 2021.
- Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

-
- Yecheng Jason Ma, Andrew Shen, Dinesh Jayaraman, and Osbert Bastani. Smodge: Versatile offline imitation learning via state occupancy matching. *arXiv preprint arXiv:2202.02433*, 2022.
- Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021.
- Liyuan Mao, Haoran Xu, Weinan Zhang, and Xianyuan Zhan. Odice: Revealing the mystery of distribution correction estimation via orthogonal-gradient update. *arXiv preprint arXiv:2402.00348*, 2024.
- Ofir Nachum and Bo Dai. Reinforcement learning via fenchel-rockafellar duality. *arXiv preprint arXiv:2001.01866*, 2020.
- Tianwei Ni, Harshit Sikchi, Yufei Wang, Tejus Gupta, Lisa Lee, and Ben Eysenbach. f-irl: Inverse reinforcement learning via state marginal matching. In *Conference on Robot Learning*, pp. 529–551. PMLR, 2021.
- Dean A Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural computation*, 3(1):88–97, 1991.
- Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3(6):233–242, 1999.
- Harshit Sikchi, Akanksha Saran, Wonjoon Goo, and Scott Niekum. A ranking game for imitation learning. *arXiv preprint arXiv:2202.03481*, 2022.
- Harshit Sikchi, Rohan Chitnis, Ahmed Touati, Alborz Geramifard, Amy Zhang, and Scott Niekum. Score models for offline goal-conditioned reinforcement learning. *arXiv preprint arXiv:2311.02013*, 2023a.
- Harshit Sikchi, Qinqing Zheng, Amy Zhang, and Scott Niekum. Dual rl: Unification and new methods for reinforcement and imitation learning. In *Sixteenth European Workshop on Reinforcement Learning*, 2023b.
- Wen Sun, Anirudh Vemula, Byron Boots, and Drew Bagnell. Provably efficient imitation learning from observation alone. In *International conference on machine learning*, pp. 6036–6045. PMLR, 2019.
- Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- Gokul Swamy, Sanjiban Choudhury, J Andrew Bagnell, and Steven Wu. Of moments and matching: A game-theoretic framework for closing the imitation gap. In *International Conference on Machine Learning*, pp. 10022–10032. PMLR, 2021.
- Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.
- Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018a.
- Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2018b.
- Faraz Torabi, Garrett Warnell, and Peter Stone. Adversarial imitation learning from state-only demonstrations. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2229–2231, 2019.

Stefan Vogt and Roland Thomaschke. From visuo-motor interactions to imitation learning: behavioural and brain imaging studies. *Journal of sports sciences*, 25(5):497–517, 2007.

Chao Yang, Xiaojian Ma, Wenbing Huang, Fuchun Sun, Huaping Liu, Junzhou Huang, and Chuang Gan. Imitation learning from observations by minimizing inverse dynamics disagreement. *arXiv preprint arXiv:1910.04417*, 2019.

Zhuangdi Zhu, Kaixiang Lin, Bo Dai, and Jiayu Zhou. Off-policy imitation learning from observations. *Advances in Neural Information Processing Systems*, 33:12402–12413, 2020.

Konrad Zolna, Alexander Novikov, Ksenia Konyushkova, Caglar Gulcehre, Ziyu Wang, Yusuf Aytar, Misha Denil, Nando de Freitas, and Scott Reed. Offline learning from demonstrations and unlabeled experience. *arXiv preprint arXiv:2011.13885*, 2020.

Appendix

6.1 Theory

6.1.1 Derivation for Action-free distribution matching

Theorem 4.1. *The dual problem to the primal occupancy matching objective (Equation 5) is given by the DILO objective in Equation 6. Moreover, as strong duality holds from Slater’s conditions the primal and dual share the same optimal solution d^* for any offline transition distribution ρ and any choice of mixture distribution ratio β .*

We start with the primal objective that matches distributions between the agent’s visitation $d(s, s', a')$ and expert’s visitation $d^E(s, s', a')$. As before ρ denotes the visitation distribution of offline data.

$$\min_{\pi} \mathcal{D}_f(\text{Mix}_{\beta}(d^{\pi}(s, s', a'), \rho) \| \text{Mix}_{\beta}(d^E(s, s', a'), \rho)), \quad (8)$$

where for any two distributions μ_1 and μ_2 , $\text{Mix}_{\beta}(\mu_1, \mu_2)$ denotes the mixture distribution with coefficient $\beta \in (0, 1]$ defined as $\text{Mix}_{\beta}(\mu_1, \mu_2) = \beta\mu_1 + (1 - \beta)\mu_2$.

Formulating the objective as a constrained objective in agent’s visitation distribution d allows us to create a primal objective that is a convex program. This is crucial in creating a dual objective that is unconstrained and easy to optimize.

$$\begin{aligned} & \max_{d \geq 0} -\mathcal{D}_f(\text{Mix}_{\beta}(d, \rho) \| \text{Mix}_{\beta}(d^E, \rho)) \\ \text{s.t. } & \sum_{a''} d(s', s'', a'') = (1 - \gamma)d_0(s', s'') + \gamma \sum_{s, a' \in \mathcal{S} \times \mathcal{A}} d(s, s', a') p(s'' | s', a'), \quad \forall s', s'' \in \mathcal{S} \times \mathcal{S}. \end{aligned} \quad (9)$$

where the constraints above dictate the constraints that any valid visitation distribution $d(s', s'')$ needs to satisfy and are our proposed modifications to the commonly known *bellman flow constraints*.

Below we outline the derivation of how these specific constraints with the mixture distribution matching objective allows us to create a dual objective that is independent of expert’s actions.

Applying Lagrangian duality and convex conjugate (??) to the above distribution matching objective, we can convert it to an unconstrained problem with dual variables $Q(s, s')$ defined for all $s, s' \in \mathcal{S} \times \mathcal{S}$:

$$\begin{aligned}
& \max_{d \geq 0} \min_{Q(s', s'')} -D_f(\text{Mix}_\beta(d, \rho)(s, s', a') \parallel \text{Mix}_\beta(d^E, \rho)(s, s', a')) \\
& + \sum_{s', s''} Q(s', s'') \left((1 - \gamma) d_0(s', s'') + \gamma \sum_{s, a'} d(s, s', a') p(s'' | s', a') - \sum_a d(s', s'', a'') \right) \quad (10) \\
& = \max_{d \geq 0} \min_{Q(s, s')} (1 - \gamma) \mathbb{E}_{d_0(s, s')} [Q(s, s')] + \mathbb{E}_{s, s', a' \sim d} \left[\gamma \sum_{s''} p(s'' | s', a') Q(s', s'') - Q(s, s') \right] \\
& - D_f(\text{Mix}_\beta(d, \rho)(s, s', a') \parallel \text{Mix}_\beta(d^E, \rho)(s, s', a')) \quad (11)
\end{aligned}$$

where the last equation uses a change of variable from s', s'' to s, s' without loss of generality.

$$\begin{aligned}
& = \max_{d \geq 0} \min_{Q(s, s')} \beta (1 - \gamma) \mathbb{E}_{d_0(s, s')} [Q(s, s')] \\
& + \beta \mathbb{E}_{s, s', a' \sim d} \left[\gamma \sum_{s''} p(s'' | s', a') Q(s', s'') - Q(s, s') \right] \\
& + (1 - \beta) \mathbb{E}_{s, s', a' \sim \rho} \left[\gamma \sum_{s''} p(s'' | s', a') Q(s', s'') - Q(s, s') \right] \\
& - (1 - \beta) \mathbb{E}_{s, a, g \sim \rho} \left[\gamma \sum_{s'} p(s'' | s', a') Q(s', s'') - Q(s, s') \right] \\
& - D_f(\text{Mix}_\beta(d, \rho)(s, s', a') \parallel \text{Mix}_\beta(d^E, \rho)(s, s', a')) \quad (12)
\end{aligned}$$

Now using the fact that strong duality holds in this problem we can swap the inner max and min resulting in:

$$\begin{aligned}
& = \min_{Q(s, s')} \max_{\text{Mix}_\beta(d, \rho)(s, s', a') \geq 0} \beta (1 - \gamma) \mathbb{E}_{d_0(s, s')} [Q(s, s')] \\
& + \beta \mathbb{E}_{s, s', a' \sim d} \left[\gamma \sum_{s''} p(s'' | s', a') Q(s', s'') - Q(s, s') \right] \\
& + (1 - \beta) \mathbb{E}_{s, s', a' \sim \rho} \left[\gamma \sum_{s''} p(s'' | s', a') Q(s', s'') - Q(s, s') \right] \\
& - (1 - \beta) \mathbb{E}_{s, s', a' \sim \rho} \left[\gamma \sum_{s''} p(s'' | s', a') Q(s', s'') - Q(s, s') \right] \quad (13)
\end{aligned}$$

$$- D_f(\text{Mix}_\beta(d, \rho)(s, s', a') \parallel \text{Mix}_\beta(d^E, \rho)(s, s', a')) \quad (14)$$

In the following derivation, we will show that the inner maximization in Eq 13 has a closed form solution even when adhering to the non-negativity constraints. Let $y(s, s', a') = \mathbb{E}_{s'' \sim p(s', a')} [Q(s', s'')] - Q(s, s')$.

$$\begin{aligned}
& \max_{\text{Mix}_\beta(d, \rho)(s, s', a') \geq 0} \mathbb{E}_{s, s', a' \sim \text{Mix}_\beta(d, \rho)(s, s', a')} \left[\gamma \sum_{s''} p(s'' | s', a') Q(s', s'') - Q(s, s') \right] \\
& - D_f(\text{Mix}_\beta(d, \rho)(s, s', a') \parallel \text{Mix}_\beta(d^E, \rho)(s, s', a'))
\end{aligned}$$

Now to solve this constrained optimization problem we create the Lagrangian dual and study the KKT (Karush–Kuhn–Tucker) conditions. Let $w(s, s', a') \triangleq \frac{\text{Mix}_\beta(d, \rho)(s, s', a')}{\text{Mix}_\beta(d^E, \rho)(s, s', a')}$, then the constraint $\text{Mix}_\beta(d, \rho)(s, s', a') \geq 0$ holds if and only if $w(s, s', a') \geq 0 \quad \forall s, s', a'$.

$$\begin{aligned} \max_{w(s,s',a')} \max_{\lambda \geq 0} & \mathbb{E}_{s,s',a' \sim \text{Mix}_\beta(d^E, \rho)(s,s',a')} [w(s,s',a')y(s,s',a')] - \mathbb{E}_{\text{Mix}_\beta(d^E, \rho)(s,s',a')} [f(w(s,s',a'))] \\ & + \sum_{s,s',a'} \lambda(w(s,s',a') - 0) \end{aligned} \quad (15)$$

Since strong duality holds, we can use the KKT constraints to find the solutions $w^*(s,s',a')$ and $\lambda^*(s,s',a')$.

1. **Primal feasibility:** $w^*(s,s',a') \geq 0 \quad \forall s,s',a'$
2. **Dual feasibility:** $\lambda^* \geq 0 \quad \forall s,s',a'$
3. **Stationarity:** $\text{Mix}_\beta(d^E, \rho)(s,s',a')(-f'(w^*(s,s',a')) + y(s,s',a') + \lambda^*(s,s',a')) = 0 \quad \forall s,s',a'$
4. **Complementary Slackness:** $(w^*(s,s',a') - 0)\lambda^*(s,s',a') = 0 \quad \forall s,s',a'$

Using stationarity we have the following:

$$f'(w^*(s,s',a')) = y(s,s',a') + \lambda^*(s,s',a') \quad \forall s,s',a' \quad (16)$$

Now using complementary slackness, only two cases are possible $w^*(s,s',a') \geq 0$ or $\lambda^*(s,s',a') \geq 0$. Combining both cases we arrive at the following solution for this constrained optimization:

$$w^*(s,s',a') = \max(0, f'^{-1}(y(s,s',a'))) \quad (17)$$

Using the optimal closed-form solution (w^*) for the inner optimization in Eq. (13) we obtain

$$\begin{aligned} \min_{Q(s,s')} & \beta(1 - \gamma)\mathbb{E}_{d_0(s,s')} [Q(s,s')] \\ & + \mathbb{E}_{s,s',a' \sim \text{Mix}_\beta(d^E, \rho)(s,s',a')} [\max(0, (f')^{-1}(y(s,s',a')))] y(s,s',a') - \alpha f(\max(0, (f')^{-1}(y(s,s',a')))) \\ & - (1 - \beta)\mathbb{E}_{s,a \sim \rho} \left[\gamma \sum_{s'} p(s'|s,a) Q(s',s'') - Q(s,s') \right] \end{aligned} \quad (18)$$

For deterministic dynamics, this reduces to the following simplified objective:

$$\begin{aligned} \min_{Q(s,s')} & \beta(1 - \gamma)\mathbb{E}_{d_0(s,s')} [Q(s,s')] \\ & + \mathbb{E}_{s,s',a' \sim \text{Mix}_\beta(d^E, \rho)(s,s',a')} [\max(0, (f')^{-1}(y(s,s',a')))] y(s,s',a') - f(\max(0, (f')^{-1}(y(s,s',a')))) \\ & - (1 - \beta)\mathbb{E}_{s,a \sim \rho} [\gamma Q(s',s'') - Q(s,s')] \end{aligned} \quad (19)$$

where $y(s,a,g) = \gamma Q(s',s'') - Q(s,s')$.

6.1.2 Analytical form of f_p^* for χ^2 divergence

For χ^2 divergence, the generator function $f(x) = (x - 1)^2$. $f'(x) = 2(x - 1)$ and correspondingly $f'^{-1}(x) = \frac{x}{2} + 1$. Substituting $f'^{-1}(x)$ in definition of f_p^* :

$$f_p^*(x) = \max(0, f'^{-1}(x))(x) - f(\max(0, f'^{-1}(x))) \quad (20)$$

Since x we substitute takes the form of residual $\text{residual} = \gamma \mathbb{E}_{s'' \sim p(\cdot|s',a')} [Q(s',s'')] - Q(s,s')$, the below pseudocode shows the implementation of f_p^* for DIL0.

```

1 def f_star_p(self, residual, type='chi_square'):
2     if type=='chi_square':
3         omega_star = torch.max(residual / 2 + 1, torch.zeros_like(residual))
4         return residual * omega_star - (omega_star - 1)**2

```

6.2 Experimental Analysis

6.3 Implementation

The algorithm for DIL0 can be found in Algorithm 1. We base the DIL0 implementation on the official implementation of pytorch-IQL <https://github.com/gwthomas/IQL-PyTorch/tree/main> that is based on IQL (Kostrikov et al., 2021). We keep the same network architecture as the original code and do not vary it across environments.

6.3.1 State-only Imitation Learning

Environments: For the offline imitation learning experiments we focus on 10 locomotion and manipulation environments from the MuJoCo physics engine (Todorov et al., 2012) comprising of Hopper, Walker2d, HalfCheetah, Ant, Kitchen, Pen, Door, Hammer, and Relocate to make a total of 24 datasets. The MuJoCo environments used in this work are [licensed under CC BY 4.0](#) and the datasets used from D4RL are also [licensed under Apache 2.0](#).

Suboptimal Datasets: We use the offline imitation learning benchmark from Sikchi et al. (2023b) that utilizes offline datasets consisting of environment interactions from the D4RL framework (Fu et al., 2020). Specifically, suboptimal datasets are constructed following the composition protocol introduced in SMODICE (Ma et al., 2022). The suboptimal datasets, denoted as 'random+expert', 'random+few-expert', 'medium+expert', and 'medium+few-expert' combine expert trajectories with low-quality trajectories obtained from the "random-v2" and "medium-v2" datasets, respectively. For locomotion tasks, the 'random/medium+expert' dataset contains a mixture of some number of expert trajectories (≤ 200) and ≈ 1 million transitions from the "x" dataset. The 'x+few-expert' dataset is similar to 'x+expert,' but with only 30 expert trajectories included. For manipulation environments we consider only 30 expert trajectories mixed with the complete 'x' dataset of transitions obtained from D4RL.

Expert Observation Dataset: To enable imitation learning from observation, we use 1 expert observation trajectory obtained from the "expert-v2" dataset for each respective environment.

Baselines: To benchmark and analyze the performance of our proposed methods for offline imitation learning with suboptimal data, we consider different representative baselines in this work: BC (Pomerleau, 1991), SMODICE (Ma et al., 2022), RCE (Eysenbach et al., 2021), ORIL (Zolna et al., 2020), IQLearn (Garg et al., 2021), ReCOIL (Sikchi et al., 2023b). SMODICE has been shown to be competitive (Ma et al., 2022) to DEMODICE (Kim et al., 2022) and hence we exclude it from comparison. SMODICE is an imitation learning method based on the dual framework, that optimizes an upper bound to the true imitation objective. ORIL adapts generative adversarial imitation learning (GAIL) (Ho & Ermon, 2016) algorithm to the offline setting, employing an offline RL algorithm for policy optimization. The RCE baseline combines RCE, an online example-based RL method proposed by Eysenbach et al. (2021). RCE also uses a recursive discriminator to test the proximity of the policy visitations to successful examples. (Eysenbach et al., 2021), with TD3-BC (Fujimoto & Gu, 2021). Both ORIL and RCE utilize a state-based discriminator similar to SMODICE, and TD3-BC serves as the offline RL algorithm. All the compared approaches only have access to the expert state-action trajectory.

The open-source implementations of the baselines SMODICE, RCE, and ORIL provided by the authors (Ma et al., 2022) are employed in our experiments. We use the hyperparameters provided by the authors, which are consistent with those used in the original SMODICE paper (Ma et al., 2022), for all the MuJoCo locomotion and manipulation environments.

In our set of environments, we keep the same hyper-parameters across tasks - locomotion, adroit manipulation, and kitchen manipulation. We train until convergence for all algorithms including baselines and we found the following timesteps to be sufficient for different set of environments: Kitchen: 1e6, Few-expert-locomotion: 500k, Locomotion: 300k, Manipulation: 500k

We keep a constant batch size of 1024 across all environments. For all tasks, we average mean returns over 10 evaluation trajectories and 7 random seeds. Full hyper-parameters we used for experiments are given in Table 3. For policy update, using Value Weighted Regression, we use the temperature τ to be 3 for all environments.

Hyperparameters for our proposed off-policy imitation learning method DIL0 are shown in Table 3.

Hyperparameter	Value
Policy learning rate	3e-4
Value learning rate	3e-4
f -divergence	χ^2
max-clip (Value clipping for policy learning)	100
MLP layers	(256,256)
β (mixture ratio)	0.5
η (orthogonal gradient descent)	0.5
τ (policy temperature)	3

Table 3: Hyperparameters for DIL0 in imitation from proprioceptive observations.

6.3.2 LfO with Image Observations

We use robomimic [Mandlekar et al. \(2021\)](#) for our imitation with image observations experiments. The following two environments are used here (the description is taken from their paper and written here for conciseness):

Lift: Object observations (10-dim) consist of the absolute cube position and cube quaternion (7-dim), and the cube position relative to the robot end effector (3-dim). The cube pose is randomized at the start of each episode with a random z-rotation in a small square region at the center of the table.

Can Object observations (14-dim) consist of the absolute can position and quaternion (7-dim), and the can position and quaternion relative to the robot end effector (7-dim). The can pose is randomized at the start of each episode with a random z-rotation anywhere inside the left bin.

Robomimic provides three datasets and two modalities of observation (Proprioceptive, Images) for both environments above. The datasets are denoted by - MH (Multi-human), MG (Machine Generated), PH(Proficient-human). We use the MG and MH datasets as the suboptimal datasets in our task and PH as the source of expert observations. MH and MG datasets consists of 200 trajectories of usually suboptimal nature and we use 50 observation-only trajectory from PH datasets. This tasks is complex by the fact that expert-level actions are mostly unseen in the suboptimal dataset and the agent needs to learn the best actions that matches expert visitation from the suboptimal dataset. We implement all algorithms in the Robomimic codebase without any change in network architecture, data-preprocessing or learning hyperparameters. We tune algorithm specific hyperparameters in a course grid for BCO, SMODICE, and DIL0 to compare the best performance of methods independent of hyperparameters. For BCO, we tune the inverse dynamics model learning epochs between [1,5,10]. For SMODICE, we tuned discriminator learning epochs between [1,5], and gradient penalty between [1,5,10,20]. To control overestimation due to learning with offline datasets in DIL0 we consider a

linear weighting α between the optimism and pessimism terms in Eq 6 as follows:

$$\begin{aligned} \text{DIL0: } \min_Q (1 - \lambda)\beta(1 - \gamma)\mathbb{E}_{\tilde{d}_0}[Q(s, s')] + \lambda\mathbb{E}_{s, s', a' \sim \text{Mix}_{\beta}(\tilde{d}^E, \rho)}[f_p^*(\gamma\mathbb{E}_{s'' \sim p(\cdot|s', a')}[Q(s', s'')] - Q(s, s'))] \\ - \lambda(1 - \beta)\mathbb{E}_{s, s', a' \sim \rho}[\gamma\mathbb{E}_{s'' \sim p(\cdot|s', a')}[Q(s', s'')] - Q(s, s')], \end{aligned} \quad (21)$$

The hyperparameters used for DIL0 can be found in Table 4. For the architecture specific hyperpa-

Hyperparameter	Value
max-clip (Value clipping for policy learning)	100
α (pessimism parameter)	0.7
β (mixture ratio)	0.5
η (orthogonal gradient descent)	0.5
τ (policy temperature)	3

Table 4: Hyperparameters for DIL0 in imitation from image observations.

rameters we refer the readers to (Mandlekar et al., 2021).